ORACLE
Health

Patient Population EHI Extract for Soarian Document Management

Patient Population EHI extract for Soarian Document Management is provided by a set of pre-defined extraction services. This documentation describes the data that is extracted via this process and the format of the data. This documentation only pertains to the output produced via the pre-defined extraction service.

Documents that are extracted as part of the pre-defined process

Documents from the extraction were filed in one of these Soarian Document Management folders. The folder that documents were originally filed to is not specified anywhere in the extract output. However, the indexes provided for a document are based on what folder the document had originally been in.

1) Enrollee
2) Medical Record
3) Encounter
4) Claim

Document output Information

Soarian Document Management systems can contain massive amounts of data. It is not unusual for a system to contain 50-100 million documents or more. This volume of documents can amount to 25 to 50 TB of data or more. The volume of any given system can vary significantly from these numbers.

When dealing with this volume of data, there will be coordination at the time of the extraction to determine how the physical data will be provided. These extractions will be customer specific implementations and will differ in terms of how the data is extracted and transferred. Therefore, it is not possible to provide an exact description of the type of media used or the number of media devices used. The receiver of the data from the original extract would have this detail.

The documentation focuses on the format of the data and the structure of the data on the media and not the specific distribution method or media used for an individual customer extraction.

Folder structure for extracted documents:

The documents are extracted to the following directory structure:

extract_(year)

Images

EXTRACTS

0

1

etc.

EXPORTS

0

1

etc.

The top level extract is segmented by year. Documents are grouped by year based on the document date.

The EXTRACTS level contains documents extracted in native format.

The EXPORTS level contains documents converted to PDF as part of the extraction.

Under both the EXTRACTS and EXPORTS levels, a new directory with sequential numbers is created to hold sets of 10,000 documents per directory. The documents are stored at this level. Each file stored in these directories will be referenced by a line in a Document Index file.

Note that the 10,000 documents count does not equate to 10,000 files in the EXTRACTS directory structure. Documents in the extracted section can have multiple files per document, each file typically representing a page.

File naming for extracted documents:

The documents are named with the following convention:

Docid_Extractid_1_Pageno.Format

Where

Docid is the unique document ID from the Soarian Document Management System

Extractid is a unique value assigned for a specific extraction run

Pageno is the page number of the file for the document. This will always be 1 for files in the EXPORTS directory structure. For files in the EXTRACTS directory structure, it will be the page number for each page file for a document.

Format is the 3-character MIME format code of the file. For files in the EXPORTS directory structure, this will always be PDF. For files in the EXTRACTS directory structure, it will be the format for the document stored.

Folder structure for Document Index files:

An index file is created for every 10,000 documents extracted. The index files are stored in the following directory path:

extract_[year]

File naming convention for Document Index Files:

The document index files are named as follows:

delimited_patient_index_[batch_num].txt

where batch_num is a sequential number starting at 1.

Note: While both the EXTRACTS and EXPORTS directory structures both are capped at 10,000 documents each and the index file is also capped at 10,000 documents each, there is no technical connection between the two. The documents referenced in the document index file can contain references to documents in the EXTRACTS and documents in the EXPORTS directories. You cannot assume there is any alignment between the documents referenced in an index file and the documents filed in any particular sub-directory.

Document formats extracted in their native format (EXTRACTS Directory):

The below list are document formats for which the data is returned in its existing format. In other words, if the document is stored as an MS Word .DOC document, it will be extracted as a .DOC document with no manipulation.

When a document is extracted in its existing format, each page/object of the document will be extracted as a separate file. There will be no manipulation or combining of pages. For some formats, the document's pages may all be stored in an object so you would get the entire document as one file (for example, a Word document). But for scanned documents, typically each page is a separate file and that is what is returned on the extract.

List of formats extracted natively:

- BMP
- CSV
- DOC
- GIF
- HTM
- JPG
- PDF
- PNG
- PPT
- RTF
- SNP
- TIF
- TXT
- XLS

Documents that are converted to PDF (EXPORTS directory):

1) Any document that is stored in a format NOT in the list in the previous section
2) Any document with annotations or signatures regardless of format
3) Any document with an overlay template associated with it regardless of format
4) Documents that are in mixed format (Ex. One file/page is TIF, one file/page is PDF)

Extract Report

In each extract_YEAR folder, there is a summary report for the extracts of documents for that year. This report is for informational purposes only. Its content does not relate to the interpretation of the data that has been extracted.

The report is named [year]_doc_types.md.html. It contains four sections:

1. Extracted Documents By DocType
2. Documents that Could Not Be Indexed by DocType
    a. This section will be blank in the final output data.
3. Documents that Could Not Be Extracted by DocType
    a. A count by doc type of all documents that were not extracted.
4. Document Extract Failure Messages
    a. This section is a table listing all documents that cannot be extracted with the following fields;
        i. Doc ID
        ii. Doc Type
        iii. Detailed Failure Message

Document index Information

Document index information or metadata is provided in an associated delimited file.

The metadata file will be pipe delimited and each record will be separated by a carriage return/line feed. The first line of the file is a header line containing the field names for the rest of the file.

Each metadata file will contain index details for up to 10,000 documents.

For a document that is extracted as native and has more than one page, there will be an individual line in the index file for each page. The pages within a document can be grouped into a document by the doc_id field in the document metadata section.

ORACLE
Health

Metadata for documents

Field Name	Description	Folder Types
hhrr	The hospital region code of the facility the patient visited.	Medical Record, Encounter, Claim
corp_id	The enrollee id of the patient.	All Folders
patient_name	The name of the patient - Last, First Middle.	All Folders
patient_birth_date	The birth date of the patient.	All Folders
gender	The patient's gender.	All Folders
mrn	The patient's medical record number.	Medical Record, Encounter, Claim
claim_no	The patient's claim number.	Claim
encounter_no	The patient's visit number.	Encounter
admit_date	The start date of the patient's visit.	Encounter
discharge_date	The date the patient's visit ended.	Encounter
hosp_svc	Hospital service performed for the patient.	Encounter
fin_class	The encounter's financial class.	Encounter
vip_indicator	The encounter's VIP indicator value.	Encounter
unit_number	The unit number. This field is deprecated.	N/A
unit_date	The unit date. This field is deprecated.	N/A
security_item	The VIP security token assigned to the folder.	All Folders

Document Metadata

All metadata files will contain the following fields:

Field Name	Description
doc_id	Unique Soarian DM document id.
version	Version number of the document.
document_type_name	Name of the document type.
document_type_desc	Description of the document type.
document_labels	Any labels attached to the document.
doc_date	Document date of the document.
create_date	Create date of the document.
modify_date	Modify date of the document.
create_user	The Soarian DM user that created the document.
file_path	Path to the physical file of the document.

- Each label identified for a document in *Soarian Document Management* will be added to index file – each label sub-delimited by a tilde (~)
- If a document is marked as ERRONEOUS in *Soarian Document Management*, an ERRONEOUS label will be created